frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 10 May 2024
DOI 10.3389/fgene.2024.1399810

OPEN ACCESS

EDITED BY
Chunhou Zheng,
Anhui University, China

REVIEWED BY
Jin-Xing Liu,
University of Health and Rehabilitation Sciences, China
Yijie Ding,
University of Electronic Science and Technology of China, China

*CORRESPONDENCE
Chang-Qing Yu,
✉ xaycq@163.com
Li-Ping Li,
✉ cs2bioinformatics@gmail.com

RECEIVED 13 March 2024
ACCEPTED 03 April 2024
PUBLISHED 10 May 2024

CITATION
Zhao Y-X, Yu C-Q, Li L-P, Wang D-W, Song H-F and Wei Y (2024), BJLD-CMI: a predictive circRNA-miRNA interactions model combining multi-angle feature information.
Front. Genet. 15:1399810.
doi: 10.3389/fgene.2024.1399810

BJLD-CMI: a predictive circRNA-miRNA interactions model combining multi-angle feature information

Yi-Xin Zhao[1], Chang-Qing Yu[1]*, Li-Ping Li[1,2]*, Deng-Wu Wang[1], Hui-Fan Song[1] and Yu Wei[1]

[1]School of information Engineering, Xijing University, Xi'an, China, [2]College of Grassland and Environment Sciences, Xinjiang Agricultural University, Ürümqi, China

Increasing research findings suggest that circular RNA (circRNA) exerts a crucial function in the pathogenesis of complex human diseases by binding to miRNA. Identifying their potential interactions is of paramount importance for the diagnosis and treatment of diseases. However, long cycles, small scales, and time-consuming processes characterize previous biological wet experiments. Consequently, the use of an efficient computational model to forecast the interactions between circRNA and miRNA is gradually becoming mainstream. In this study, we present a new prediction model named BJLD-CMI. The model extracts circRNA sequence features and miRNA sequence features by applying Jaccard and Bert's method and organically integrates them to obtain CMI attribute features, and then uses the graph embedding method Line to extract CMI behavioral features based on the known circRNA-miRNA correlation graph information. And then we predict the potential circRNA-miRNA interactions by fusing the multi-angle feature information such as attribute and behavior through Autoencoder in Autoencoder Networks. BJLD-CMI attained 94.95% and 90.69% of the area under the ROC curve on the CMI-9589 and CMI-9905 datasets. When compared with existing models, the results indicate that BJLD-CMI exhibits the best overall competence. During the case study experiment, we conducted a PubMed literature search to confirm that out of the top 10 predicted CMIs, seven pairs did indeed exist. These results suggest that BJLD-CMI is an effective method for predicting interactions between circRNAs and miRNAs. It provides a valuable candidate for biological wet experiments and can reduce the burden of researchers.

KEYWORDS

circRNA-miRNA interaction, circRNA, miRNA, graph embedding, autoencoder

1 Introduction

Circular RNAs (circRNAs) are a new class of endogenous non-coding RNAs with covalently closed loops that are important components of gene transcription. In comparison to traditional linear RNA, circRNA, due to it is end-to-end covalent closure and the absence of a 5′ cap or 3′ poly(A) tail (Zheng et al., 2020), is less susceptible to degradation by exonucleases, rendering it is structure more stable. CircRNA was first discovered in virus-infected plant particles in 1976. However, due to it is low expression levels and sparse occurrence, it was initially considered a byproduct of gene transcriptional splicing errors or "splicing noise," receiving limited attention from researchers at that time. Hsu and Coca-Prados (1979) used electronic microscopy in 1979 to provide indications of the existence of circRNA in the cytoplasm of eukaryotic cell lines. With the maturity of high-throughput sequencing technology as well as the continuous development of bioinformatics, researchers have discovered that circRNA is abundant in eukaryotes and performs a crucial role in various biological processes. In 2013, Memczak et al. (2013) demonstrated through sequence analysis that circRNA has important regulatory functions. In the same year, Hansen et al. (2013) discovered a highly expressed circular RNA that binds with miR-7 within human and mouse brain. Additionally, they identified a testis-specific circRNA acting as a miR-138 sponge. This led to the inference that the miRNA sponge effect formed by circRNA is a widespread phenomenon. Simultaneously, various biological functions of circRNA have been gradually understood by humans. For instance, circRNA can act as a scaffold for protein complex assembly, regulate gene expression, and modulate selective splicing RNA-protein interactions (Deng et al., 2019). As understanding of circRNA and miRNA deepens, an increasing amount of research evidence suggests the existence of connections between the two. CircRNAs participate in organic processes, and their dysregulation and mutations can affect disease progression. As an example, Gong et al. (2023) discovered that hsa_circ_0064644 could inhibit the proliferation and migration of osteosarcoma cells by acting as a miR-424-5p sponge to regulate eIF4B and YRDC. Wang et al. (2022a) demonstrated that hsa_circ_0005505 modulates KIF2A expression by acting as a miR-603 sponge, and silencing hsa_circ_0005505 will cause self-apoptosis of breast cancer cells, which cannot normally multiply, move and invade other tissues *in vitro*. Causing tumor growth in the body to slow down. Pan et al. (Pan and Binghua, 2023) confirmed that hsa_circ_0135761 positively regulates EFR3 by competitively binding to miR-654-3p. Reducing the gene level of hsa_circ_0135761 promotes apoptosis in NP carcinoma cells, as well as inhibits the increase and relocation of nasopharyngeal carcinoma cells. Therefore, circRNA may serve as a potential biomarker. Examining the potential correlation between circRNAs and miRNAs holds significant clinical guidance for biologists in diagnosing and treating diseases.

Before the popularity of computational models, identifying the relation of circRNAs to miRNAs typically relied on classical biological experimental methods. However, validating these experimental results often proved to be quite cumbersome. With the discovery of an increasing number of circular RNAs, the growing number of miRNAs to be validated has posed significant challenges to traditional biological experimental verification methods. As computational models gradually address the drawbacks of traditional biological experiments, including long experimental cycles, high costs, small-scale studies, and susceptibility to external interference, an increasing number of researchers have begun to predict the interrelationships between circRNAs and miRNAs with the help of computational modeling. This alleviates the burden of experimental validation and provides researchers with a broader perspective. In 2018, Li et al. (2018) explored data integration principles using a machine learning approach to analyze a variety of downstream tasks using a computer-based perspective. In 2019, the computational framework CMASG was introduced by Qian et al. (2021b), which utilizes singular value decomposition and graph variational autoencoder to extract linear and nonlinear features from circRNA and miRNA. Additionally, they integrated the framework to predict the interactions between them. In 2021, Lan et al. (2021) suggested a new method named NECNA for network-based embedding. This method utilizes GIP kernel similarity networks of circRNA and miRNA, along with their associated networks, to construct a heterogeneous network. Through neighborhood regularized matrix factorization and inner product, NECNA predicts the interaction between circRNA and miRNA.

As methods for extracting and fusing features continue to mature, a single feature cannot interpret all the information in an organism, so researchers have begun to combine multiple features to interpret information in organisms. In 2021, Qian et al. (2021a) proposed a model, MKSVM, which fuses multiple feature information extracted from protein sequences through a central kernel alignment-based multiple kernel learning (MKL-CKA) algorithm to predict DBP. In 2022, a computational model called WSCD was introduced by Guo et al. (2022). It utilizes graph embedding and word embedding to extract features and integrates convolutional neural networks (CNN) and deep neural networks (DNN) to deduce the potential interactions between circRNA and miRNA. In the same year, Qian et al. (2022a) proposed a prediction model for adverse drug reactions, which constructs two spatial RBMs to predict drug-side effect associations by fusing the similarity feature matrix of the drug chemical structure information and the similarity feature matrix of the association mapping based on the central kernel alignment (MKL-CKA) algorithm, as well as the adjacency matrix supplemented by Weighted K nearest known neighbors. Yu et al. (2022) introduced a model named SGCNCMI. This model employs a graph neural network with a contribution mechanism to aggregate multi-modal information from nodes for predicting CMIs. Qian et al. (2022b) proposed a model, MvKSRC, which combines multi-view features such as amino acid composition, evolutionary information and amino acid physicochemical information to further predict membrane protein types by Kernel Sparse Representation based Classification (KSRC). Wang et al. (2022b) designed a computational method, KGDCMI, based on the fusion of multiple sources of information to predict interactions between circRNA and miRNA. This method combines sequence and similarity to obtain attribute features, combined with the behavior features, the extracted feature vectors are sent to the deep neural network for prediction. In 2023, Li et al. (2023) introduced a multi-source information fusion model, DeepCMI. This model integrates various information, including sequence similarity matrices and Gaussian interaction kernel features, to construct multi-source features. Through linear embedding prediction of CMI by enhanced feature extraction through linear embedding.

Although the existing models mentioned above have achieved relatively influential prediction results, they still inevitably have certain limitations in terms of efficiency and methodology, and many prediction models are built on statistical models and machine learning algorithms that lack an understanding of biological information. Consequently, the prediction results may be difficult to interpret or unreliable. To address the above issues, we proposed a novel computational model-based approach called BJLD-CMI for predicting

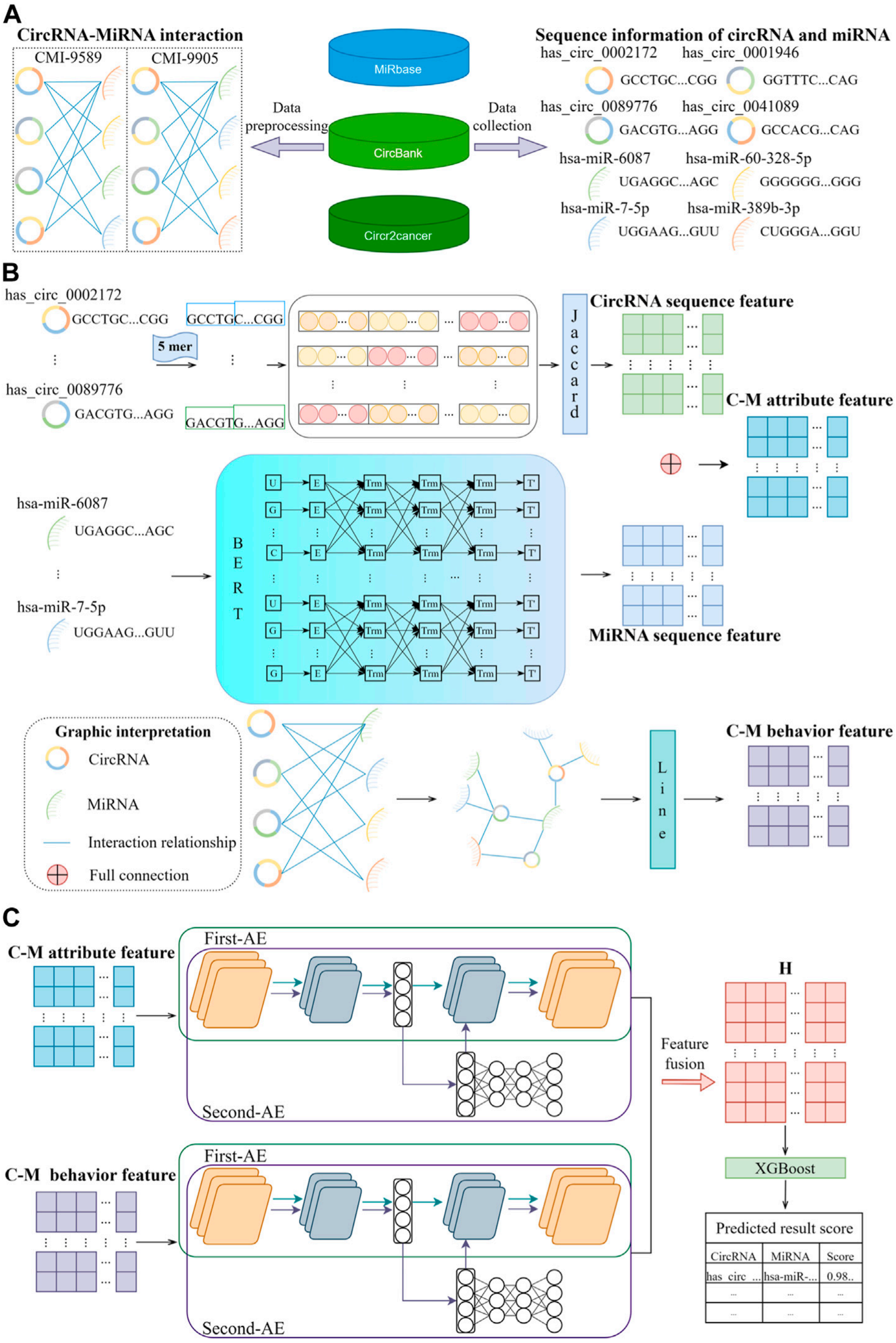

FIGURE 1
BJLD-CMI workflow diagram. **(A)** Data were collected and cleaned from MiRbase, CircBank, and CircR2cancer databases to obtain the CMI-9589 dataset and CMI-9905 dataset, respectively. **(B)** Jaccard, Bert, and Line were applied to extract the known attribute features and behavioral features of CMI. **(C)** Use Autoencoder in Autoencoder Networks to fuse the attribute features and behavioral features, and the fused features are predicted and analyzed by the XGBoost classifier for each CMI.

circRNA-miRNA interactions in this study. Firstly, We utilized the Jaccard and Bert (Devlin et al., 2018) methods to extract features from the circRNA and miRNA sequences respectively and organically integrate them into attribute features of CMI. Secondly, we employed the graph embedding method Line (Tang et al., 2015) to extract behavioral features of CMI based on the graph information of circRNA-miRNA interactions. We then introduce the Autoencoder in Autoencoder Networks (Zhang et al., 2019) model to fuse the behavioral and attribute features of circRNA and miRNA, obtaining comprehensive features between them. Finally, an XGBoost (Chen and Guestrin, 2016) classifier is utilized to predict potential CMIs. We conducted a comprehensive evaluation of the model performance based on five-fold cross-validation (5-fold CV). In the validation on the CMI-9905 dataset, we achieved remarkable performance, with the AUC

TABLE 1 CMI data set information used in BJLD.

Dataset	CircRNA	MiRNA	Interaction
CMI-9589 (CircBank)	2,115	821	9,585
CMI-9905 (CircBank, Circr2cancer)	2,346	962	9,905

reaching 90.69%, accuracy as high as 88.36%, and precision reaching 85.31%. We also compared different classifiers and obtained favorable results. Additionally, we conducted a case study on BJLD-CMI, wherein we validated the top 10 predicted CMI pairs from the experiment through the latest literature in the PubMed database. It was found that 7 of these pairs have already been confirmed to have a relationship. Based on these experimental results, we conclude that the BJLD-CMI model plays a significant role in predicting the interaction between circRNA and miRNA, providing effective guidance for biological experiments to identify circRNA as relevant miRNA sponges. Figure 1 illustrates the workflow of BJLD-CMI.

2 Materials and methods

2.1 Dataset

For this study, to assess the model's ability to predict CMI, we utilized two commonly used datasets, namely, the CMI-9589 dataset and the CMI-9905 dataset. The CMI-9589 dataset is sourced from the CircBank (Liu et al., 2019) database, which is a comprehensive human circRNA database containing detailed annotations for 140,790 human circRNAs from various sources. In addition to providing basic information about circRNA, CircBank also offers a set of interaction data between circRNA and miRNA for predicting and analyzing miRNA interactions. Based on the cleaning and summarizing the data, we acquired known 9,589 circRNA-miRNA pairs, among which 2,115 circRNAs and 821 miRNAs were involved. The CircR2Cancer (Lan et al., 2020) database is a manually curated database that associates circRNA with cancer. The CMI-9905 dataset, obtained by Wang et al. (2022b), comprises data on circRNA-miRNA interactions from the public database CircR2Cancer, including combined 318 circRNA-miRNA pairs among 238 circRNAs and 230 miRNAs. By integrating the data of the two databases, 9,905 good-quality CMI pairs were final acquired, comprising 2,346 circRNAs and 962 miRNAs. In this study, we primarily utilized the CMI-9905 dataset and regarded it as the positive sample, detailed information is available in Table 1. Subsequently, we randomly selected 9,905 unknown CMI pairs from the data pool of $2346 \times 962 - 9905 = 2,246,947$ as negative samples.

2.2 Constructing attribute characteristics

In bioinformatics, researchers typically analyze the nucleotide sequences of RNA to extract features such as nucleotide composition, base pair frequency, sequence length, etc. These features help reveal the structure, function, and relationships with other biomolecules of RNA. Due to the substantial differences in the length of circRNA nucleotide sequences, with some long circRNAs containing thousands of nucleotides and short circRNAs containing only a few hundred nucleotides, we chose to use the Jaccard similarity coefficient to extract attribute features from circRNA sequences. This is because the Jaccard similarity coefficient can better reflect the similarity between sequences of different lengths. miRNAs usually have relatively short lengths, typically between 20 and 22 nucleotides. To extract attribute features from miRNA sequences, we employed the Bert model. Finally, we integrate the attribute features of circRNAs and miRNAs depending on the interaction relationships and finally obtain the CMI attribute features of known relationship pairs.

2.2.1 Jaccard similarity coefficient

The Jaccard model is one of the fundamental models for similarity recognition, while the Jaccard similarity coefficient is an important metric used to measure the similarity between two sets. The factor takes values between 0 and 1, and the closer the value is to 1, the more similar the two sets are. The Jaccard similarity coefficient, widely employed in bioinformatics, serves to assess the dissimilarity or similarity between finite sets of samples (Wang et al., 2021). In order to extract the most effective sequence information, we use a moving window whose window size is 5 and whose stride size is 1 to split the sequence. This divides a circRNA of length L into sets of length $L/5$. Then, we utilize the Jaccard similarity coefficient to gauge the representation of differences in circRNA C_a and other circRNA sequences in the sample, represented as Formula 1:

$$J_{C_a} = \left| \frac{\boldsymbol{C_a} \cap \boldsymbol{C_e}}{\boldsymbol{C_a} \cup \boldsymbol{C_e}} \right| \tag{1}$$

Where $\boldsymbol{C_e}$ represents the sequence set of 2,346 different circRNAs, with $e \in [1, 2346]$.

2.2.2 Natural language processing model of bert

Bert (Devlin et al., 2018) is an unsupervised pre-trained language model used for natural language processing tasks. It consists a bidirectional multi-layer Transformer encoder stack. Bert learns from two unsupervised pre-training tasks, namely, Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks. In the MLM task, model learning anticipates some tokens of the input sequence by shadowing them randomly. In the NSP task, model learns to judge whether two sentences were adjacent or not. As shown in Figure 2, Bert preprocesses the input information, represented as Formula 2 when the input is miRNA sequence information:

$$\boldsymbol{input_{pre}} = ([\boldsymbol{CLS}], \boldsymbol{U}, \boldsymbol{G}, \ldots, [\boldsymbol{SEP}], \boldsymbol{A}, \boldsymbol{G}, \ldots, \boldsymbol{C}, [\boldsymbol{SEP}]) \tag{2}$$

$[CLS]$ is a special token added at the beginning of each input sequence, and $[SEP]$ is a special token used as a separator. Bert's input is the sum of Token embedding, Segment embedding, and Position embedding, represented as Formula 3:

$$\begin{aligned}\boldsymbol{input} = \boldsymbol{Embedding_{Tok}}(\boldsymbol{input_{pre}}) + \boldsymbol{Embedding_{Seg}}(\boldsymbol{input_{pre}}) \\ + \boldsymbol{Embedding_{Pos}}(\boldsymbol{input_{pre}})\end{aligned} \tag{3}$$

At the encoding layer, Bert comprises a 12-layer Transformer encoding network, with each layer having a hidden size of 768, aimed at extracting latent features and establishing correlations

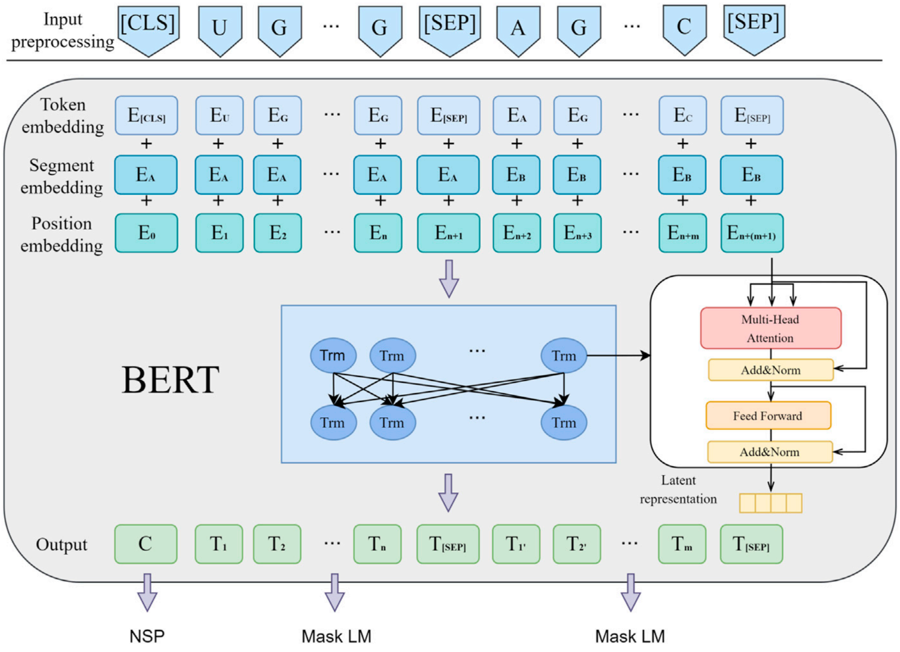

FIGURE 2
Basic flowchart of Bert pre-training.

between contexts. Compared to ELMo (Sarzynska-Wawer et al., 2021), Bert uses Transformer blocks as extractors, pre-trained through MLM (Masked Language Model) to enhance semantic feature extraction capabilities. In contrast to GPT (Radford et al., 2018), Bert switches from unidirectional to bidirectional encoding, leveraging all context information for each word and predicting and reconstructing the original data from corrupted input data through autoencoding. Compared to unidirectional encoders that can only utilize leading information for semantic information extraction, Bert exhibits stronger semantic information extraction capabilities.

In this study, we obtained a set of 9,905 relationship pairs between circRNA and miRNA, involving the IDs of 962 miRNAs. To study these miRNAs in depth, we introduced that Bert model. By querying miRbase (Griffiths-Jones et al., 2007), we retrieved sequence information for the 962 miRNAs corresponding to their IDs. This sequence information was used as input, and after processing through the Bert model, we obtained a digitized representation of the output, extracting attribute features representing the miRNAs.

2.3 Graph embedding for behavioral feature extraction

Graph embedding (Yan et al., 2006) is a technique that maps every node in a diagram structure to a low-dimensional vector space, and it plays a crucial role in many graph data analysis tasks. In the field of bioinformatics, graph embedding is widely used to study complex biological relationships (Yi et al., 2022), such as molecular interactions and gene regulatory networks. Graph embedding utilizes known interactions among circRNAs with miRNAs to obtain a matrix that the behavior feature of circRNAs with miRNAs. An interaction information network is regarded as $G = (V, E)$, in which V denotes the collection of vertices (data objects) and E denotes the collection of edges interactions between vertices. Most graph embedding methods only consider first-order proximity, such as Deepwalk (Perozzi et al., 2014). In this study, we employ the Line (Tang et al., 2015) graph embedding algorithm, which preserves both first-order and second-order proximity, thus better preserving the global structure of the network. By learning low-dimensional representations of nodes, we can more effectively capture the similarity and interaction of nodes in the graph structure.

The first-order proximity in the interaction network of circRNA and miRNA is the local pairwise proximity between two vertices, V_c and V_m. If there is no edge, the first-order proximity is 0. For each undirected edge (c, m), the formula for the first-order joint probability distribution of vertices V_c and V_m is defined as Formula 4:

$$p_1(V_c, V_m) = \frac{1}{1 + exp\left(-\vec{u}_c^{T} \cdot \vec{u}_m\right)} \tag{4}$$

Where $\vec{u}_c, \vec{u}_c \in R^d$ are the low-dimensional representation vectors of vertices V_c and V_m. It is first-order goal function is shown in Formula 5:

$$O_1 = \sum_{(c,m)\in E} W_{cm} \cdot log\, p_1(V_c, V_m) \tag{5}$$

Where W_{cm} is the connection weight between vertices V_c and V_m. The second-order proximity in the interaction network of CircRNA and miRNA is the similarity between the neighborhood network structures of two vertices V_c and V_m. If there is no neighborhood network structure, the second-order proximity is 0. In the second-order proximity, each vertex has two tasks: Task 1: the vertex itself, Task 2: a specific context with other vertices. For each edge (c, m), the second-order joint probability distribution formula is Formula 6:

$$p_2(V_m|V_c) = \frac{exp\left(\vec{u}_m^{'T} \cdot \vec{u}_c\right)}{\sum_{k=1}^{|V|} exp\left(\vec{u}_k^{'T} \cdot \vec{u}_c\right)} \tag{6}$$

where $|V|$ is the number of vertices or contexts. It is second-order goal function is shown in Formula 7:

$$O_2 = \sum_{(c,m)\in E} W_{cm} \cdot log\, p_2((V_m|V_c)) \tag{7}$$

To expedite the learning process, we modify the Formula 7 using negative sampling. The modified objective function is expressed as Formula 8:

$$O_2 = log\, \sigma\left(\vec{u}_m^{'T} \cdot \vec{u}_c\right) + \sum_{i=1}^{K} E_{V_n \sim P_n(V)}\left[log\, \sigma\left(-\vec{u}_n^{'T} \cdot \vec{u}_c\right)\right] \tag{8}$$

Where $\sigma(x) = \frac{1}{(1+exp(-x))}$ is the sigmoid function, K is the number of negative samples, and $P_n(V) \sim d(V)^{\frac{3}{4}}$ is usually set, where $d(V)$ represents the degree of vertex V.

2.4 A neural network model for double layer nested automatic encoder

Autoencoder is an unsupervised learning neural network model designed to learn efficient representations of data. It can be utilized for tasks such as data compression, denoising, and feature extraction. The autoencoder model is instrumental in aiding the exploration and prediction of interactions among non-coding RNAs. In this study, we use the Autoencoder in Autoencoder Networks (Zhang et al., 2019) model to effectively fuse circRNA and miRNA with multi-angle features, that is, to fuse the two angular features of circRNA and miRNA, so that the fused features not only have complementarity between behavioral features and attribute features, but also have complementarity between behavioral features and attribute features. It also has the consistency between behavior characteristics and attribute characteristics. The Autoencoder in Autoencoder Networks model mainly includes the First-AE network and the Second-AE network, which can learn single-angle feature representation and complete multi-angle feature representation together. Then the First AE network is used to extract the implicit information of each Angle automatically, and the degradation process of the Second AE network is used to encode the implicit information of each Angle into the potential representation. We represent the sample of multi-angle features as $X = X^{(1)}, X^{(2)}, ..., X^{(V)}, X^{(V)} \in R^{d_v \times n}$ is the feature matrix of the $V - th$ Angle feature, where V represents the number of angles of the $V - th$ Angle feature, n represents the number of samples of the $V - th$ Angle feature, and d_v represents the feature dimension of the $V - th$ Angle feature.

2.4.1 First-AE network

We use $f(X^{(v)}; \xi_{ae}^{(v)})$ to denote the First-AE network for the $V - th$ angle feature, where $\xi_{ae}^{(v)} = \{W_{ae}^{(c,v)}, b_{ae}^{(c,v)}\}_{c=1}^{C}$ is the parameter set for all layers, $W_{ae}^{(c,v)}$ represents the relevant weights for the $c - th$ layer, $b_{ae}^{(c,v)}$ represents the relevant biases for the $c - th$ layer, and C represents the number of layers for nonlinear transformations. The first $C/2$ encoding layers encode the input feature vector into a new vector, and the last $C/2$ decoding layers reconstruct the new vector. When the input feature vector is $x_i^{(v)} = z_i^{(0,v)} \in R^{d_v}$, the output of the $c - th$ layer is Formula 9:

$$z_i^{(c,v)} = \sigma\left(W_{ae}^{(c,v)} z_i^{(c-1,v)} + b_{ae}^{(c,v)}\right), c = 1, 2, \ldots, C \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid activation function, and when the input is the feature matrix $X^{(v)} = [x_1^{(v)}, x_2^{(v)}, \ldots, x_n^{(v)}] \in R^{d_v \times n}$ of the $v - th$ Angle feature, the corresponding reconstruction formula is as Formula 10:

$$Z^{(C,v)} = \left[z_1^{(c,v)}, z_2^{(c,v)}, \ldots, z_n^{(c,v)}\right] \tag{10}$$

Where $Z^{(C,v)}$ is the reconstructed representation of the i sample in the $v - th$ Angle feature. We obtain a low-dimensional representation $Z^{(\frac{C}{2},v)}$ through minimal reconstruction loss, the minimal reconstruction loss is obtained as Formula 11:

$$\min_{\{\xi_{ae}^{(v)}\}_{v=1}^{V}} \frac{\sum_{v=1}^{V} \left\|X^{(v)} - Z^{(C,v)}\right\|_F^2}{2} \tag{11}$$

We encode the obtained low-dimensional angle feature representation $Z^{(\frac{C}{2},v)}$ into a holistic latent information H for the entire angle feature, where $H \in R^{k \times n}$ and k represents the complete spatial dimension.

2.4.2 Second-AE network

The degradation reduction network of the Second-AE network uses a fully connected neural network (FCNN) to realize that each angular feature can be represented by a new common representation of the whole. We use $g(\mathrm{H}; \xi_{dr}^{(v)})$ to represent the degradation restoration network of the $v - th$ angle feature, where $\xi_{ae}^{(v)} = \left\{W_{dr}^{(s,v)}, b_{dr}^{(s,v)}\right\}_{s=1}^{S}$, and $S + 1$ is the number of layers in the degradation restoration network. We take $H = G^{(0,v)}$ as input, then $G^{(s,v)} = [g_1^{(s,v)}, g_2^{(s,v)}, \ldots, g_n^{(s,v)}]$, where $g_i^{(s,v)} = \sigma(W_{dr}^{(s)} g_i^{(s-1,v)} + b_{dr}^{(s,v)})$, and the formula for the goal of degradation reduction is Formula 12:

$$\min_{\{\xi_{dr}^{(v)}\}_{v=1}^{V}} \frac{\sum_{v=1}^{V} \left\|Z^{(\frac{C}{2},v)} - G^{(S,v)}\right\|_F^2}{2} \tag{12}$$

2.4.3 Coupling the First-AE network with the Second-AE network

In the same framework, we learned new vector representations for each angle feature (via the First-AE network) and latent representations for the complete multi-angle features (via the Second-AE network) by coupling the First-AE network with the Second-AE network. The objective function of Autoencoder in Autoencoder Networks model is summarized as Formula 13:

$$\min_{\{\xi_{ae}^{(v)}, \xi_{dr}^{(v)}\}_{v=1}^{V}, H} \frac{\sum_{v=1}^{V} \left[\left\|X^{(v)} - Z^{(C,v)}\right\|_F^2 + \lambda \left\|Z^{(\frac{C}{2},v)} - G^{(S,v)}\right\|_F^2\right]}{2} \tag{13}$$

Here, λ represents the balance between the consistency and complementarity of multi-angle features.

2.5 XGBoost classifier

XGBoost (Chen and Guestrin, 2016) is referred to as extreme gradient boosting, and it is an integrated learning algorithm based on gradient boosting decision trees (GBDT). It is employed to solve machine learning issues such as classification, regression, and ranking. XGBoost benefits from its efficiency, regularization processing, feature importance analysis, and ability to handle missing values, making it a powerful tool for many data science problems. We employ XGBoost as a classifier for predicting circRNA-miRNA interactions. Through multiple iterations, we gradually construct a decision tree model, emphasizing error samples to enhance model performance. The objective function of XGBoost is composed of the loss function and regularization, and is formulated as Formula 14:

$$\boldsymbol{L}^{(u)} = \sum_{i=1}^{M} \boldsymbol{l}\left(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i^{(u-1)} + \boldsymbol{f}_u(\boldsymbol{x}_i)\right) + \sum_{u} \boldsymbol{\varphi}(\boldsymbol{f}_u) \tag{14}$$

Here, i represents the $i-th$ sample, u represents the $u-th$ tree, $\boldsymbol{y}_i$ is the true value of the $i-th$ sample x_i, $\hat{y}_i$ is the predicted value of the $i-th$ sample x_i, l is the differentiable loss function computing the difference between the predicted value $\hat{y}_i$ and the target value y_i, and $\varphi(\cdot)$ represents the complexity of the tree. By expanding the second-order Taylor series and regularization term, separately optimizing the loss function term and regularization term, and merging similar terms, the final objective function is obtained as Formula 15:

$$\boldsymbol{L}^{(u)} = \sum_{j=1}^{U} \left[\boldsymbol{B}_n \boldsymbol{w}_n + \frac{(\boldsymbol{D}_n + \boldsymbol{\lambda})\boldsymbol{w}_n^2}{2} \right] + \boldsymbol{\gamma U} \tag{15}$$

Here, B_n and D_n respectively represent the sums of the first and second-order partial derivatives of the samples contained in leaf node n, w_n represents the weight of the $n-th$ leaf node, and U represents the number of trees.

3 Results

3.1 Evaluation indicators criteria

Cross-validation is a commonly used method in the field of machine learning for assessing model performance and reducing the bias of evaluation results. In this work, we employ 5-fold cross-validation (5-fold CV) to assess the predictive power of BJLD-CMI over CMI-9905 dataset. We initially randomly divided the dataset into five subsets, ensuring a balanced distribution of categories in each subset as much as possible. Four subsets were utilized for pieces of training of the model and then one remaining subset was used for validation of the model. This process is repeated five times, ensuring that each subset is used for validation once. The results of the five validations were averaged to get the final performance evaluation metrics (Wang et al., 2018). The experimental evaluation of BJLD-CMI includes accuracy (ACC), precision (Prec.), recall (Rec.), F1-score (F1), and Matthews correlation coefficient (MCC) as reliability assessment criteria. The formula for accuracy is Formula 16, for precision is Formula 17, for recall is Formula 18, for F1-score is Formula 19, and for Matthews correlation coefficient is Formula 20:

$$\boldsymbol{ACC.} = \frac{\boldsymbol{TN} + \boldsymbol{TP}}{\boldsymbol{TN} + \boldsymbol{TP} + \boldsymbol{FN} + \boldsymbol{FP}} \tag{16}$$

$$\boldsymbol{Prec.} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{17}$$

$$\boldsymbol{Rec.} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{18}$$

$$\boldsymbol{F1 - score} = \frac{\boldsymbol{2} \times (\boldsymbol{Precision} \times \boldsymbol{Recall})}{\boldsymbol{Precision} + \boldsymbol{Recall}} \tag{19}$$

$$\boldsymbol{MCC.} = \frac{(\boldsymbol{TP} \times \boldsymbol{TN}) - (\boldsymbol{FP} \times \boldsymbol{FN})}{\sqrt{(\boldsymbol{TP} + \boldsymbol{FP}) \times (\boldsymbol{TP} + \boldsymbol{FN}) \times (\boldsymbol{TN} + \boldsymbol{FP}) \times (\boldsymbol{TN} + \boldsymbol{FN})}} \tag{20}$$

In the above formula, true positives (TP) indicate the sample counts in which the model predicts that circRNAs interact with miRNAs and in which the interaction is realistically confirmed, false positives (FP) indicate the sample counts in which the model predicts that circRNAs interact with miRNAs but in reality are not interaction, true negative (TN) indicates the sample count in which the model predicts that circRNAs not interact with miRNAs and realistically confirms that there is no interaction, false negative (FN) indicates the sample count in which the model predicts that circRNAs not interact with miRNAs but the interaction is confirmed in reality. We also plotted the Receiver Operating Characteristic (ROC) curve of BJLD-CMI under 5-fold cross-validation (Zweig and Campbell, 1993) and computed the AUC to evaluate the performance of the models.

3.2 Evaluation model prediction ability

In this experiment, we tested the performance of BJLD-CMI in predicting circRNA-miRNA interactions over the CMI-9905 dataset with a 5-fold cross-validation method. Table 2 lists the details of the experimental results. From Table 2, we can see that our model obtained a mean accuracy of 83.41%. The accuracies for the five experiments were 83.47%, 84.07%, 82.53%, 83.06%, and 83.90% respectively, with a standard deviation of 0.56%. In the evaluation criteria of precision (Prec), recall (Rec), F1-score, and Matthews correlation coefficient (MCC), BJLD-CMI demonstrated an accuracy of 85.31%, 80.70%, 82.94%, and 66.91%, with respective standard deviations of 0.38%, 0.93%, 0.64%, and 1.10%. In addition, we also computed the AUC and AUPR that BJLD-CMI generated over the CMI-9905 dataset with their ROC curves and PR curves plotted. Concerning AUC, the five experiments yielded results of 90.56%, 91.41%, 90.21%, 90.49%, and 90.75%, with a mean value of 90.69% and a standard deviation of 0.45%. Concerning AUPR, the five experiments resulted in 88.87%, 90.06%, 89.01%, 88.66%, and 89.39% with a mean value of 89.20% and a standard deviation of 0.55%, respectively. Figure 3A shows the ROC curves generated by five experiments, and Figure 3B shows the PR curves generated by five experiments. Through the experimental results described above, it is clearly observed that the BJLD-CMI is able to predict CMI effectively on CMI-9905, and shows excellent comprehensive performance, exhibits good application prospects, and may be considered as a potential tool for exploring the unknown CMI.

3.3 Evaluation comparison of different dimensions of line

The model utilizes the Line algorithm in graph embedding to learn low-dimensional embeddings of nodes in the circRNA-

TABLE 2 Cross-validation results of BJLD-CMI on the CMI-9905 dataset.

Test set	ACC. (%)	Prec. (%)	Rec. (%)	F1-score (%)	MCC. (%)	AUC (%)
1	83.47	85.12	81.12	83.07	67.01	90.56
2	84.07	85.94	81.47	83.65	68.24	91.41
3	82.53	84.93	79.10	81.91	65.22	90.21
4	83.06	85.03	80.26	82.58	66.23	90.49
5	83.90	85.55	81.57	83.51	67.87	90.75
Average	83.41 ± 0.56	85.31 ± 0.38	80.70 ± 0.93	82.94 ± 0.64	66.91 ± 1.10	90.69 ± 0.45

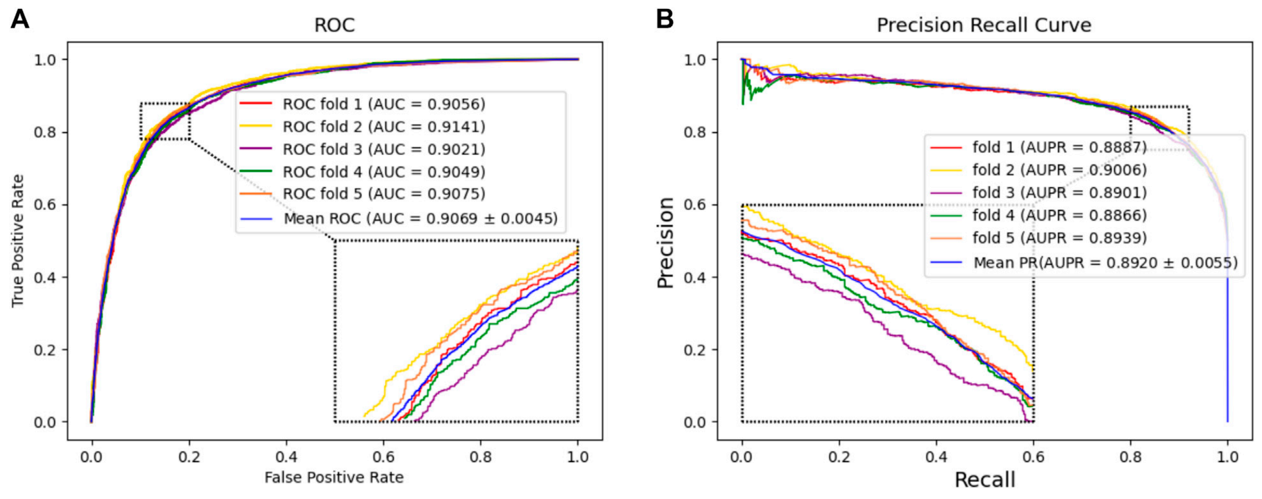

FIGURE 3
Performance evaluation of BJLD-CMI **(A)** ROC curve obtained by BJLD-CMI on the CMI-9905 dataset **(B)** PR curve obtained by BJLD-CMI on the CMI-9905 dataset.

TABLE 3 Results of 5-fold cross-validation on CMI-9905 dataset with different Line dimensions.

Dimensions	Mean ACC	Mean Prec	Mean Rec	Mean F1-score	Mean MCC	Mean AUC
32	0.7675	0.7637	0.7748	0.7692	0.5351	0.8449
64	0.8151	0.8232	0.8026	0.8127	0.6303	0.8858
128	0.8341	0.8531	0.8070	0.8294	0.6691	0.9069
256	0.8029	0.8094	0.7923	0.8008	0.6059	0.8869
512	0.7473	0.7867	0.6786	0.7287	0.4994	0.8431

miRNA relationship network, preserving the network structure and relationships between nodes. Determining the dimension is a crucial factor in describing the features of circRNA and miRNA. Choosing a big dimension could increase the computational workload and intricacy of the suggested model, resulting in longer execution times and potentially lower accuracy. Conversely, selecting a small dimension might lead to an insufficient feature extraction. Therefore, we chose a series of commonly used dimensions, specifically 32, 64, 128, 256, 512, and conducted a 5-fold cross-validation experiment to select the optimal dimension. Based on the experimental results in Table 3; Figure 4, we observed a continuous improvement in the overall performance of the model with increasing dimensions. When the dimension parameter increases to 128, the model achieves optimal performance, as reflected in the maximum values of ACC, AUC, Prec., and MCC. However, when the dimension exceeds 128, the performance gradually declines. Therefore, we decided to fix the dimension parameter at 128.

3.4 Comparison of different classifiers

Our proposed BJLD-CMI model uses the XGBoost classifier for data training and classification tasks over the CMI-9905. To validate the performance of the XGBoost classifier, we replaced it with four

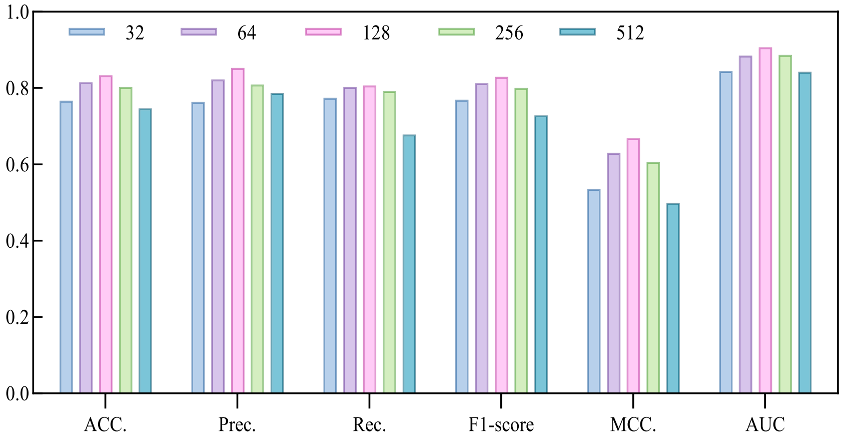

FIGURE 4
Performance bar visualization comparison with different Line dimensions on CMI-9905 dataset.

TABLE 4 Results of different traditional classifiers and XGBoost in 5-fold cross-validation on CMI-9905 dataset.

Classifier	Testing set	ACC. (%)	Prec. (%)	Rec. (%)	F1-score (%)	MCC. (%)	AUC (%)
XGBoost	Average	83.41	85.31	80.70	82.94	66.91	90.69
	SD	0.56	0.38	0.93	0.64	1.10	0.45
RF	Average	82.30	85.14	78.26	81.56	64.82	89.77
	SD	0.30	0.35	0.75	0.39	0.58	0.36
GaussianNB	Average	77.40	84.24	67.40	74.88	55.92	87.68
	SD	0.78	0.65	1.51	1.05	1.46	0.54
SVM	Average	77.36	83.09	68.69	75.21	55.55	81.44
	SD	0.64	0.88	1.02	0.76	1.28	0.80
LR	Average	77.30	81.67	70.41	75.62	55.13	81.30
	SD	0.19	0.30	0.33	0.22	0.39	0.46

other classifiers while keeping the dataset and other conditions unchanged. These classifiers are Random Forest (RF) (Breiman, 2001), Gaussian Naive Bayes (GaussianNB), Support Vector Machine (SVM) (Suykens and Vandewalle, 1999), and Logistic Regression (LR) (Dreiseitl and Ohno-Machado, 2002). The performance of these five classifiers in predicting CMI was evaluated by 5-fold cross-validation and their categorization performance is compared. Table 4 concludes the average results of the five classifiers combined with the CMI-9905 dataset after 5-fold cross-validation and is presented in line graph form. From Figure 5, we can intuitively observe that XGBoost achieved the highest results in six evaluation metrics, including ACC, Prec, Rec, F1, MCC, and AUC. This indicates that XGBoost performs better in predicting unknown CMI in the proposed model.

3.5 Comparison of the existing method

Currently, numerous outstanding computational approaches have been proposed, relying on benchmark datasets CMI-9589 [from the Circbank database (Liu et al., 2019)] and CMI-9905 [from the Circbank database (Liu et al., 2019) and Circr2cancer database (Lan et al., 2020)]. These methods include WSCD (Guo et al., 2022), SGCNCMI (Yu et al., 2022), KGDCMI (Wang et al., 2022b), DeepCMI (Li et al., 2023), CMIVGSD (Qian et al., 2021b), GCNCMI (He et al., 2022), JSNDCMI (Wang et al., 2023), aim to forecast potential CMIs. In order to further assess BJLD-CMI's predictive performance, we have compared it to these methods in two datasets, respectively. For fairness, we chose the AUC generated by the fivefold CV method as the parameter for evaluation. Table 5 presents the contrasting outcomes of CMIVGSD, SGCNCMI, KGDCMI, GCNCM, JSNDCMI, and DeepCMI with BJLD-CMI utilizing the CMI-9589 dataset. Table 6 shows the contrasting outcomes of KGDCMI, WSCD, SGCNCM, JSNDCMI, and DeepCMI with BCMCMI utilizing the CMI-9905 dataset. The results in Tables 5, 6 indicate that our model achieved the highest AUC results, surpassing the averages by 0.00986 and 0.03158, respectively. Overall, BJLD-CMI demonstrates strong competitiveness among existing methods.

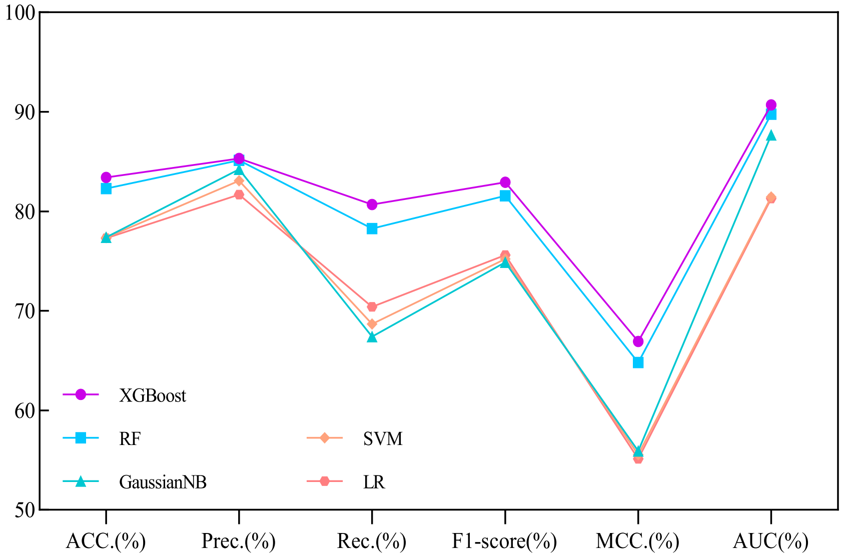

FIGURE 5
Performance comparison line chart of different traditional classifiers and XGBoost on the CMI-9905 dataset.

TABLE 5 AUC values of related models and BJLD-CMI on 5-fold cross-validation in CMI-9589 dataset.

Models	CMIVGSD	SGCNCMI	KGDCMI	GCNCMI	JSNDCMI	DeepCMI	BJLD-CMI
AUC	0.8804	0.9015	0.9041	0.9320	0.9415	0.9480	0.9495
AUPR	0.8629	0.9011	0.8937	0.9396	0.9403	0.9416	0.9474

TABLE 6 AUC values of related models and BJLD-CMI on 5-fold cross-validation in CMI-9905 dataset.

Models	KGDCMI	WSCD	SGCNCMI	JSNDCMI	DeepCMI	BJLD-CMI
AUC	0.8930	0.8923	0.8942	0.9003	0.9054	0.9069
AUPR	0.8767	0.8935	0.8887	0.8999	0.8978	0.8920

3.6 Case studies

In order to validate the genuine predictive ability of BJLD-CMI for miRNA-associated circRNA, we performed a case study utilizing the CMI-9905 dataset. In the experiment, we trained the BJLD-CMI model using known CMIs extracted from the CMI-9905 dataset and then used the trained model to predict unknown CMIs. Following the acquisition of the prediction outcomes, we organized the prediction scores in descending order and validated the top 10 circRNA-miRNA pairs in the published literature. The specific findings are outlined in Table 7. From the table, we can conclude that 7 pairs have been confirmed in PubMed, confirming the involvement of circRNA as miRNA sponges in the biological processes of diseases such as lung cancer (Wang et al., 2019; Yao et al., 2019; Zhou et al., 2019), prostate cancer (Wu et al., 2019), and gastric cancer (Zhong et al., 2018; Liang et al., 2019). It's worth noting that the lack of confirmation in existing literature for the other 4 pairs does not necessarily negate the possibility of an interaction between them. The results of the case study indicate that BJLD-CMI is a powerful tool with the prospect of exploring the interaction of unknown circRNAs with miRNAs.

4 Conclusion

With the popularity of computational models and the booming development of bioinformatics, people are gradually realizing the importance of the associative relationships between circRNAs and miRNAs in various biological processes as well as in the treatment of diseases. By applying computational models to predict CMI, we can get a deeper understanding of the unrevealed hidden networks between circRNAs and miRNAs, and thus study their roles in regulating gene expression and participating in organic processes. Exploring the correlation between circRNAs and miRNAs provides biologists with new ideas, which are important clinical guidance for the diagnosis and treatment of diseases.

On this work, we suggest a computationally grounded model, BJLD-CMI, to forecast circRNA with miRNA interaction

TABLE 7 Top 10 CMI pairs predicted by BJLD-CMI.

Num	miRNA	circRNA	Evidence	Cancer
1	hsa-miR-1183	hsa_circ_0004015	PMID:30509491	Non-Small Cell Lung Cancer
2	hsa-miR-4667-3p	hsa_circ_0002172	Unconfirmed	Unconfirmed
3	hsa-miR-135a-5p	hsa_circ_0001946	PMID:30841451	Lung Adenocarcinoma
4	hsa-miR-181c-5p	hsa_circ_0001427	PMID:30674872	Prostate Cancer
5	hsa-miR-139-3p	hsa_circ_0000592	PMID:31189743	Gastric Cancer
6	hsa-miR-638	hsa_circ_0000177	PMID:30010402	Glioma
7	hsa-miR-1224-3p	hsa_circ_0001731	Unconfirmed	Unconfirmed
8	hsa-miR-214-5p	hsa_circ_0000993	PMID:30215537	Gastric Cancer
9	hsa-miR-619-5p	hsa_circ_0004939	Unconfirmed	Unconfirmed
10	hsa-miR-330-5p	hsa_circ_0001727	PMID:32010565	Non-Small Cell Lung Cancer

relationships. In this model, we first convert miRNA sequences into digital representations using natural language processing techniques, apply Jaccard similarity coefficients to obtain the feature expressions of circRNAs through the moving window method, and construct the corresponding attribute feature matrices from the known circRNA with miRNA relationship pairs. Subsequently, from the known circRNA with miRNA relationship network, we build the corresponding behavioral feature matrix using the graph embedding method Line. In addition, the Autoencoder in Autoencoder Networks model is used to learn the new vector representation of each Angle feature and the potential representation of the complete multi-angle feature respectively from the perspective of the behavior and attribute features of circRNA and miRNA, so that the obtained features not only have the complementarity between the behavior feature and the attribute feature but also have the consistency. On the CMI-9589 and CMI-9905 datasets, BJLD-CMI achieved excellent results using the XGBoost classifier. To evaluate the performance of BJLD-CMI, we conducted experiments comparing different classifiers and experiments comparing with other models. The results indicate that BJLD-CMI outperforms other models. We also conducted a case study, and among the top ten ranked circRNA-miRNA pairs in prediction scores, 7 pairs were verified in our literature search on PubMed. This provides new insights for research on diseases such as non-small cell lung cancer, lung adenocarcinoma, prostate cancer, and gastric cancer.

These results indicate that the BJLD-CMI model can predict the underlying relationship between circRNAs and miRNAs efficiently and is a reliable predictive model, but there are still some limitations. Firstly, the BJLD-CMI model is dependent on the amount of data on known circRNA-miRNA interaction, and too large a gap of positive and negative samples can have a significant impact on the correctness of the model's predictions. Secondly, different feature extraction methods and parameter settings may also impact the model's predictions. Additionally, the BJLD-CMI model could not make direct predictions for circRNA-miRNA pairs that have no known interactions. In future research, we will continue exploring the application of NLP in extracting information from biological sequence data and integrating additional perspectives of biological feature information to enhance the accuracy and reliability of the model.

Data availability statement

The datasets used in this paper can be found in the CircR2Cancer database (http://www.biobdlab.cn:8000/) and the CircBank database (http://www.circbank.cn/). The source code can be found at https://github.com/YXzhaok/BJLDCMI.

Author contributions

Y-XZ: Conceptualization, Data curation, Writing–original draft, Writing–review and editing, Methodology, Resources, Software, Validation. C-QY: Writing–original draft, Writing–review and editing, Conceptualization, Data curation, Methodology, Resources, Software, Validation. L-PL: Writing–original draft, Writing–review and editing, Conceptualization, Data curation, Methodology, Resources, Software, Validation. D-WW: Conceptualization, Writing–original draft, Data curation, Methodology, Resources, Software, Validation. H-FS: Writing–original draft. YW: Writing–original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China under Grant Nos 62273284, 62172338, and 62072378.

Acknowledgments

The authors would like to thank all editors and reviewers for their valuable suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Chen, T., and Guestrin, C. (2016). XGBoost. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min*. doi:10.1145/2939672.2939785

Deng, L., Zhang, W., Shi, Y., and Tang, Y. (2019). Fusion of multiple heterogeneous networks for predicting circRNA-disease associations. *Sci. Rep.* 9, 9605. doi:10.1038/s41598-019-45954-x

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv Prepr. arXiv:1810.04805*.

Dreiseitl, S., and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inf.* 35, 352–359. doi:10.1016/s1532-0464(03)00034-0

Gong, Z., Shen, P., Wang, H., Zhu, J., Liang, K., Wang, K., et al. (2023). A novel circular RNA circRBMS3 regulates proliferation and metastasis of osteosarcoma by targeting miR-424-eIF4B/YRDC axis. *Aging (Albany NY)* 15, 1564–1590. doi:10.18632/aging.204567

Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2007). miRBase: tools for microRNA genomics. *Nucleic Acids Res.* 36, D154–D158. doi:10.1093/nar/gkm952

Guo, L.-X., You, Z.-H., Wang, L., Yu, C.-Q., Zhao, B.-W., Ren, Z.-H., et al. (2022). A novel circRNA-miRNA association prediction model based on structural deep neural network embedding. *Briefings Bioinforma.* 23, bbac391. doi:10.1093/bib/bbac391

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi:10.1038/nature11993

He, J., Xiao, P., Chen, C., Zhu, Z., Zhang, J., and Deng, L. (2022). GCNCMI: a graph convolutional neural network approach for predicting circRNA-miRNA interactions. *Front. Genet.* 13, 959701. doi:10.3389/fgene.2022.959701

Hsu, M.-T., and Coca-Prados, M. (1979). Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature* 280, 339–340. doi:10.1038/280339a0

Lan, W., Zhu, M., Chen, Q., Chen, B., Liu, J., Li, M., et al. (2020). CircR2Cancer: a manually curated database of associations between circRNAs and cancers. *Database* 2020, baaa085. doi:10.1093/database/baaa085

Lan, W., Zhu, M., Chen, Q., Chen, J., Ye, J., Liu, J., et al. (2021). Prediction of circRNA-miRNA associations based on network embedding. *Complexity* 2021, 1–10. doi:10.1155/2021/6659695

Li, Y.-C., You, Z.-H., Yu, C.-Q., Wang, L., Hu, L., Hu, P.-W., et al. (2023). DeepCMI: a graph-based model for accurate prediction of circRNA–miRNA interactions with multiple information. *Briefings Funct. Genomics*, elad030. doi:10.1093/bfgp/elad030

Liang, M., Liu, Z., Lin, H., Shi, B., Li, M., Chen, T., et al. (2019). High-throughput sequencing reveals circular RNA hsa_circ_0000592 as a novel player in the carcinogenesis of gastric carcinoma. *Biosci. Rep.* 39. doi:10.1042/BSR20181900

Liu, M., Wang, Q., Shen, J., Yang, B. B., and Ding, X. (2019). Circbank: a comprehensive database for circRNA with standard nomenclature. *RNA Biol.* 16, 899–905. doi:10.1080/15476286.2019.1600395

Li, Y., Wu, F.-X., and Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. *Briefings Bioinforma.* 19, 325–340. doi:10.1093/bib/bbw113

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338. doi:10.1038/nature11928

Pan, J., and Binghua, X. (2023). Circular RNA EFR3A promotes nasopharyngeal carcinoma progression through modulating the miR-654-3p/EFR3A axis. *Cell. Mol. Biol.* 69, 111–117. doi:10.14715/cmb/2023.69.12.18

Perozzi, B., al-Rfou, R., and Skiena, S. (2014). "Deepwalk: online learning of social representations," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, USA, Augest-24-27, 2014 (IEEE), 701–710.

Qian, Y., Ding, Y., Zou, Q., and Guo, F. (2022a). Identification of drug-side effect association via restricted Boltzmann machines with penalized term. *Briefings Bioinforma.* 23, bbac458. doi:10.1093/bib/bbac458

Qian, Y., Ding, Y., Zou, Q., and Guo, F. (2022b). Multi-view kernel sparse representation for identification of membrane protein types. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20, 1234–1245. doi:10.1109/TCBB.2022.3191325

Qian, Y., Jiang, L., Ding, Y., Tang, J., and Guo, F. (2021a). A sequence-based multiple kernel model for identifying DNA-binding proteins. *BMC Bioinforma.* 22, 291–318. doi:10.1186/s12859-020-03875-x

Qian, Y., Zheng, J., Zhang, Z., Jiang, Y., Zhang, J., and Deng, L. (2021b). "CMIVGSD: circRNA-miRNA interaction prediction based on variational graph auto-encoder and singular value decomposition," in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), China, Dec. 8 2023 (IEEE).

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding with unsupervised learning. *Tech. rep., Technical report, OpenAI.*

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., et al. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res.* 304, 114135. doi:10.1016/j.psychres.2021.114135

Suykens, J. A., and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Process. Lett.* 9, 293–300. doi:10.1023/a:1018628609742

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). "Line," in Proceedings of the 24th International Conference on World Wide Web, USA, May 18-22, 2015 (IEEE).

Wang, X.-F., Yu, C.-Q., Li, L.-P., You, Z.-H., Huang, W.-Z., Li, Y.-C., et al. (2022b). KGDCMI: a new approach for predicting circRNA–miRNA interactions from multi-source information extraction and deep learning. *Front. Genet.* 13, 958096. doi:10.3389/fgene.2022.958096

Wang, X.-F., Yu, C.-Q., You, Z.-H., Li, L.-P., Huang, W.-Z., Ren, Z.-H., et al. (2023). A feature extraction method based on noise reduction for circRNA-miRNA interaction prediction combining multi-structure features in the association networks. *Briefings Bioinforma.* 24, bbad111. doi:10.1093/bib/bbad111

Wang, F., Li, J., Li, L., Chen, Z., Wang, N., Zhu, M., et al. (2022a). Circular RNA circ_IRAK3 contributes to tumor growth through upregulating KIF2A via adsorbing miR-603 in breast cancer. *Cancer Cell. Int.* 22, 81. doi:10.1186/s12935-022-02497-y

Wang, L., You, Z.-H., Li, J.-Q., and Huang, Y.-A. (2021). IMS-CDA: prediction of CircRNA-disease associations from the integration of multisource similarity information with deep stacked autoencoder model. *IEEE Trans. Cybern.* 51, 5522–5531. doi:10.1109/TCYB.2020.3022852

Wang, L., You, Z.-H., Yan, X., Xia, S.-X., Liu, F., Li, L.-P., et al. (2018). Using two-dimensional principal component analysis and rotation forest for prediction of protein-protein interactions. *Sci. Rep.* 8, 12874. doi:10.1038/s41598-018-30694-1

Wang, Y., Xu, R., Zhang, D., Lu, T., Yu, W., Wo, Y., et al. (2019). Circ-ZKSCAN1 regulates FAM83A expression and inactivates MAPK signaling by targeting miR-330-5p to promote non-small cell lung cancer progression. *Transl. Lung Cancer Res.* 8, 862–875. doi:10.21037/tlcr.2019.11.04

Wu, G., Sun, Y., Xiang, Z., Wang, K., Liu, B., Xiao, G., et al. (2019). Preclinical study using circular RNA 17 and micro RNA 181c-5p to suppress the enzalutamide-resistant prostate cancer progression. *Cell. Death Dis.* 10, 37. doi:10.1038/s41419-018-1048-1

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., and Lin, S. (2006). Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. pattern analysis Mach. Intell.* 29, 40–51. doi:10.1109/TPAMI.2007.12

Yao, Y., Hua, Q., Zhou, Y., and Shen, H. (2019). CircRNA has_circ_0001946 promotes cell growth in lung adenocarcinoma by regulating miR-135a-5p/SIRT1 axis and activating Wnt/β-catenin signaling pathway. *Biomed. Pharmacother.* 111, 1367–1375. doi:10.1016/j.biopha.2018.12.120

Yi, H. C., You, Z. H., Huang, D. S., and Kwoh, C. K. (2022). Graph representation learning in bioinformatics: trends, methods and applications. *Brief. Bioinform* 23, bbab340. doi:10.1093/bib/bbab340

Yu, C.-Q., Wang, X.-F., Li, L.-P., You, Z.-H., Huang, W.-Z., Li, Y.-C., et al. (2022). SGCNCMI: a new model combining multi-modal information to predict circRNA-related miRNAs, diseases and genes. *Biology* 11, 1350. doi:10.3390/biology11091350

Zhang, C., Liu, Y., and Fu, H. (2019). "Ae2-nets: autoencoder in autoencoder networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, China, June 24 2022 (IEEE), 2577–2585.

Zheng, K., You, Z.-H., Li, J.-Q., Wang, L., Guo, Z.-H., and Huang, Y.-A. (2020). iCDA-CGR: identification of circRNA-disease associations based on Chaos Game Representation. *PLOS Comput. Biol.* 16, e1007872. doi:10.1371/journal.pcbi.1007872

Zhong, S., Wang, J., Hou, J., Zhang, Q., Xu, H., Hu, J., et al. (2018). Circular RNA hsa_circ_0000993 inhibits metastasis of gastric cancer cells. *Epigenomics* 10, 1301–1313. doi:10.2217/epi-2017-0173

Zhou, Y., Zheng, X., Xu, B., Chen, L., Wang, Q., Deng, H., et al. (2019). Circular RNA hsa_circ_0004015 regulates the proliferation, invasion, and TKI drug resistance of non-small cell lung cancer by miR-1183/PDPK1 signaling pathway. *Biochem. Biophysical Res. Commun.* 508, 527–535. doi:10.1016/j.bbrc.2018.11.157

Zweig, M. H., and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* 39, 561–577. doi:10.1093/clinchem/39.4.561